ADAPTIVE PLAYOUT OF DIGITAL PACKET AUDIO WITH PACKET FORMAT INDEPENDENT JITTER REMOVAL

* * * * *

CROSS-REFERENCE TO RELATED APPLICATIONS

Not applicable.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

Not applicable.

BACKGROUND OF THE INVENTION

The present invention relates to adaptive playout of audio packets transported over a packet network without reliance on packet header time-stamp information for jitter removal. In particular, the present invention relates to effective playout through removal of jitter and minimization of delay independent of packetization format .

Voice over packet networks or VoIP requires that the voice or audio signal be packetized and then transmitted. The transmission path will typically take the packets through both packet switched and circuit switched networks between each termination of the transmission. The analog voice signal is first converted to a digital signal and compressed at a gateway connected between a terminal equipment and the packet network. The gateway produces a pulse code modulated (PCM) digital stream from the analog voice.

The PCM stream is analyzed in the gateway and processed according to the parameters of the gateway, such as echo suppression, silence detection and DTMF tone detection. Detected tones are passed separately without encoding. The voice PCM samples are passed to a CODEC for processing prior to packet assembly.

The CODEC creates voice frames from the PCM stream according to the parameters of the codec used. The creation of frames from the PCM stream typically includes compression. The frames are of known size and, based upon the specified rate, are of a determinable time duration. Each frame contains a set number of bits of the PCM stream dependant on the codec used for bi-directional conversion between analog audio and digital packets.

The frames are then assembled into packets by a packet assembler which combines a set number of sequential frames into a single packet data payload. A header, such as a real time protocol (RTP) header is attached to each packet payload to provide a sequence number for identification of the packet and a time stamp for the packet. In the case of RTP format, information about the length of the packet is provided in the IP header. The gatekeeper then assigns an IP address to the packet corresponding to the designated destination of the voice signal to which the packet belongs. An IP header is added to the packet to designate the origination and destination IP addresses for the packet. A UDP header containing source and destination sockets can also be added to the packet.

The packets are routed through the packet network based upon the IP address information. The packet may pass through several switches and routers and the signal in digital and analog form may pass through both packet switches and circuit switches respectively. The packet is likely to accumulate delay as it passes between the near and far end terminal equipment, through the near and far end gateways, through the packet and PSTN networks and through switches.

Because this accumulated delay is erratic and unpredictable and further because each packet may take a different path through the networks, delay can cause the packets to arrive out of sequence and/or with gaps or overlaps. Gapping and overlapping of packets is referred to as delay and the variance in delay from one packet to the next is called jitter. Delay and jitter are measured by comparison of the end time stamp of one packet with the start time stamp of the next packet. If the next packet is received before the end time stamp of the previous packet, there is overlapping delay. If the latest packet is received after the end time stamp of the current packet, the difference in the time is the delay gap. Conditions in the packet network can also cause the loss of packets, referred to as packet loss.

Voice packets are generated at a constant rate and represent continuous and ordered speech. Voice packets should be played out at the receiving end in the same order and at the same rate to accurately reproduce the original analog speech. Because of some inherent loss and delay in a packet network, the packets are reassembled and played out as close to the original sequence and rate as possible to achieve acceptable reproduction.

The receiving gateway will first remove the IP and UDP headers from the packets. Next the RTP information is read and the voice frames extracted from the packet. The RTP information is used to ensure that the packets are in the proper sequence. If a packet is missing, or out of order, the gateway must compensate for the missing frames in that packet in order to avoid undesirable distortion of the voice signal after frame reassembly. If one or more frames in a sequence are missing, the previous frame is repeated at a decreased volume to fill in the gap(s) left by the missing frame(s). If the missing frame subsequently arrives, too late for inclusion in the reassembled sequence of frames, the packet is discarded.

In order to compensate for jitter, the receiving gateway utilizes the sequence and time stamp of the RTP header to smooth the playout by compensating for jitter and/or packet loss by removing gaps and overlaps in the frame sequence. The receiving gateway includes a Voice Playout Unit, VPU, with a FIFO memory buffer for temporary storage of the packets. The purpose of the buffer is to remove the effect of packet arrival jitter from the voice playout. This is accomplished by adding delay before playout, such that the delay is greater than or equal to the maximum jitter encountered. The maximum delay available will be determined by the size of the buffer. Any extra delay before playout will not distort the audio but may reduce audio quality because of the addition of unnecessary delay to the system which can be noticed by users if the delay is of sufficient length. Insufficient delay will cause poor playout quality, because data will be unavailable for playout when packets are late, causing the playout to hesitate and sound distorted.

The reassembled sequence of frames is processed in a codec to return the PCM stream for playout.

SUMMARY OF THE INVENTION

In a non-adaptive, fixed playout delay, three parameters define the delay characteristics of the buffer, Figure1. The first is the minimum delay. The minimum delay is set to be sufficient to remove jitter under standard network conditions that is sufficient to hold an incoming packet while another packet is playing. The minimum delay can be set based upon the segment size of the codec. Minimum delay can for example be set to twice the time covered by one segment of the codec to compensate for anticipated network delay. For example, in a codec with a 10msc sample size, the minimum delay would be set to 20msc.

The second parameter of the buffer implementation is the nominal delay. The nominal delay is defined by the open channel message to the DSP and is used as the initial delay of the buffer. The voice playout unit will maintain the nominal delay to be equal to or greater than the minimum delay. The minimum delay will be used if the nominal delay is less than the minimum delay. In a non-adaptive voice playout environment, where no minimum delay is established, the voice playout unit will allow the delay to drop to zero and then be reset to the nominal delay.

The third parameter is the maximum delay. The maximum delay is also specified in the open channel message to the DSP. The maximum delay is determined so as to maintain acceptable playout quality. The maximum delay provides an upper bound to the delay in the buffer. If inserting a segment into the buffer will cause the buffer to exceed the maximum delay, then the segment will be discarded.

When the buffer, Figure 1, is converted to an adaptive voice playout unit, the constraints of the minimum and maximum delays are maintained. The nominal delay parameter is no longer set but instead is calculated and reset from determination of the network jitter. The system detects and measures the jitter of the received packets, Figure 2, by determining the difference between the expected receive time t_{exp} of each packet and the actual received time t_{rec} of the packet. The jitter value is averaged using a low pass filter initialized with the value of the nominal delay provided in the open channel message. The nominal delay is therefore initially set to the nominal delay defined by the user through the open channel message in the exemplary embodiment, however, the initial nominal delay can be set by any desired method. The nominal delay is recalculated from a linear function of the jitter upon the receipt of each network packet. This allows the system to improve voice quality if the jitter is high, by increasing playout delay and to eliminate unneeded delay if the jitter is low. The value of the nominal delay is always held between the minimum delay and the maximum delay, inclusive.

Playout of the received packets is begun once the nominal delay setting in the buffer is reached. The nominal delay value can be equal to the minimum delay, the maximum delay or any value in between. Playout continues and the nominal delay is computed on an ongoing real time basis, however, adjustments to the delay are not normally made during playout to avoid audio distortion which would result from adjusting delay during voice. Adjustments are only made during silence which is indicated by a SID packet. Once an SID packet is received and playout stops, the last computed nominal delay value is retained and not updated until voice restarts after the silence period. At the restart of voice, the retained nominal delay is used to initialize playout delay. This delay in implementation of a new nominal value can have the effect of lengthening or decreasing periods of silence. As SID packets can only be received when the transmitters VAD is enabled, adaptive playout will not be able to adjust the delay during normal operations if the VAD is disabled on the remote transmitter.

Optionally, a user can enable adaptive playout without the reception of a SID. After applying hysteresis to the computed delay value, voice packets can be repeated or dropped as desired or necessary. The hysteresis prevents dropping or repeating unless the new delay is much different than the old delay. This minimizes but does not eliminate distortion from dropping or repeating voice.

BRIEF DESCRIPTION OF THE DRAWINGS

For a better understanding of the nature of the present invention, reference is had to the following figures and detailed description, wherein like elements are accorded like reference numerals, and wherein:

Figure 1 is a diagram illustrating an exemplary FIFO buffer.

Figure 2 is a diagram of illustrating an exemplary packet.

Figure 3 is a diagram illustrating an exemplary series of packets with varying delay.

DETAILED DESCRIPTION OF PREFERRED EXEMPLARY EMBODIMENTS

According to the present invention, jitter is determined by noting the arrival (receive time) of a current packet and by determining the duration of a previous packet from the codec and the payload bit size of the packet. The duration of each packet should be the same, based upon the codec and protocol used by the network. Based upon the receive time t_{rec} and the duration (length l) of the current packet, the expected receive time t_{exp} of the subsequent packet can be determined.

As illustrated in Figure 1, the buffer has a predetermined maximum length, t_{max}, set by the maximum delay time. The buffer will also have a set minimum delay, t_{min}, which is at least larger than the size of a segment and is set to the length of two segments in the exemplary embodiment. The nominal delay, t_{nom}, is initially set and is reset based upon the calculation described below in reference to Figure 3.

For example, in Figure 3, the VPU notes the receive time t_{rec-1} of a the first packet, packet 1 and calculates the length l_1 of packet 1 based on the payload size in bits and the codec playout time to bit size ratio. For example a 10 byte (80bit) payload using a G729 codec with a 10ms/80bits ratio will contain 10ms of voice, a 160 bit payload using a codec with a 10ms/80bits ratio will contain 20ms of voice. The VPU will add the appropriate time (eg. 10 or 20 ms) to the receive time t_{rec-1} of the first packet to determine the expected receive time t_{exp-2} of the second packet.

The VPU next notes the arrival time t_{rec-2} of the subsequent packet, packet 2. The difference in the anticipated arrival time t_{exp-2} and the recorded arrival time t_{rec-2} of the subsequent packet, packet 2, is the calculated delay d_{1-2} between packet 1 and packet 2 and is used to determine jitter. The delay between packets 1 and 2 is a calculated delay gap because packet 2 arrived late, ie. after its anticipated arrival time. The arrival and calculated delay of subsequent packets is calculated in the same manner, as illustrated in Figure 3. For packets which arrive too early, eg. packet 4, there is an calculated delay overlap. The calculated delay of each packet is taken as the absolute value of the difference of anticipated and actual arrival time so that calculated delay is always a positive number. The calculated relative delay d for each packet, i, is maintained on a running average in real time to reset the nominal delay for the FIFO buffer when appropriate to allow for proper playout of the received data packets. Individual calculated delay for each packet, i, is continually determined by the formula: $d_i = | t_i - t_{i-1} + l_{i-1} |$

The calculated delay is the absolute value of the difference between anticipated arrival time and actual arrival time because some packets will arrive late (eg. packets 2 and 3) while some packets will arrive early (eg. packet 4). The nominal delay used to set the buffer length must be a positive number representative of the absolute value of the individual calculated delay caused by jitter in the network.

Although the exemplary embodiment illustrated detecting the delay between adjacent packets, eg. packet 1 and packet 2, the teachings of the present invention could be used to determine the delay between and two or more packets in the sequence. For example, the teachings of the present invention could be used to determine the delay between packets 1 and 3 by detecting the arrival time t_{rec-1} of packet 1 and determining the anticipated arrival time t_{exp-3} of packet 3 by adding twice the duration l of the packet to t_{rec-1}. The delay between packets 1 and 3 would be twice the average of d_{1-2} and d_{2-3}. The average delay between packets 1, 2 and 3 can therefore be determined by dividing the delay between packets 1 and 3 by two. The formula: $d_i = \mid t_i - t_{i-2} + 2l_{i-1} \mid / 2$, can be used to calculate the running average delay. Jitter determination for maintenance of the nominal delay can be based upon a running average instead of a running calculation of each individual delay, if desired. For example, the delay between packets 1 and 3 would be calculated and averaged, next the delay between packets 2 and 4 would be calculated and averaged, then the delay between packets 3 and 5 and so on.

Further, in order to reduce processing at the VPU, it is not necessary to calculate delay each time a packet is received. By using the alternative calculation method above, and performing the calculation for every second value of i, delay can be calculated only half as often as a packet is received. For example the delay between packets 1 and 3 would be calculated and averaged, then the delay between packets 3 and 5 would be calculated and averaged, next the delay between packets 5 and 7 would be calculated and averaged. It would not be necessary to calculate the other packet delays. Processing at the VPU is reduced because only the arrival time of every other packet need be noted and the calculation of delay need only be made half as often.

Further reduction in processing can be achieved by averaging over 3 or more sequential packets. The equation $d_i = | t_i - t_{i-n} + nl_{i-1} | / n$, would be used where n is the number of packets over which averaging occurs and where i is indexed by n. The balance between processor savings and jitter accuracy determination has to be made dependant upon network conditions. However, because delay is both additive and subtractive, ie some packets arrive late and have a delay gap while others arrive early with a delay overlap, on average over a sufficient number of packets, the value of the sum of the delays will be zero. The calculation of average delay will only be useful over a limited number of packets, dependant upon network conditions.

By calculating the jitter based upon arrival time and packet length, the present invention is independent of the time stamp and time stamp reference of the RTP header or other headers in the packet network. The present invention can accurately determine jitter and set the nominal delay for the VPU buffer to achieve optimized playout with minimal network delay.

Because many varying and different embodiments may be made within the scope of the inventive concept herein taught, and because many modifications may be made in the embodiments herein detailed in accordance with the descriptive requirements of the law, it is to be understood that the details herein are to be interpreted as illustrative and not in a limiting sense.